6Ш

УДК 004.934.1

В.Ю. Шелепов, А.В. Ниценко
Донецкий национальный технический университет
Институт проблем искусственного интеллекта МОН Украины и НАН Украины
Украина, 83048, м. Донецк, ул. Артема 118 б

Сегментация речевого сигнала, соответствующего заранее известному слову

A.V. Nicenko, V.Ju. Shelepov
Donetsk National Technical University
Institute of Artificial Intelligence MES of Ukraine and NAS of Ukraine, c. Donetsk
Ukraine, 83048, c. Donetsk, Artema st., 118 b

Segmentation of Speech Signal Corresponding to Beforehand Known Word

В.Ю. Шелєпов, А.В. Ніценко
Донецький національній технічний університет.
Інститут проблем штучного інтелекту МОН України і НАН України, м. Донецьк
Україна, 83048, м. Донецьк, вул. Артема 118 б

Сегментація мовного сигналу, що відповідає наперед відомому слову

В работе предлагаются алгоритмы, исправляющие и уточняющие априорную сегментацию речевого сигнала, соответствующего русскому слову в случае, когда это слово известно заранее.
Ключевые слова: априорная сегментация, условная сегментация,
широкая фонетическая классификация, метка.

The article contains algorithms for correction and clarification of segmentation of speech signal corresponding to beforehand known word.
Key words: a priory segmentation, provisional segmentation, wide phonetic classification, mark.

У роботі пропонуються алгоритми, які виправляють та уточнюють апріорну сегментацію мовного сигналу, якщо він відповідає російському слову, що відомо наперед.
Ключові слова: апріорна сегментація, умовна сегментація, широка фонетична класифікація, метка.

В работах [1], [2] описаны предложенные авторами методы сегментации речевого сигнала, то есть автоматического разбиения его на участки, отвечающие отдельным звукам русской речи, с одновременной классификацией этих участков в рамках широкой фонетической классификации (W – гласный звук, C – звонкий согласный, F – глухой фрикативный, P – глухой взрывной). Эта процедура играет важную роль в процессе распознавания как отдельно произносимых слов, так и распознавания слитной речи. В частности, на нее опирается развиваемый авторами метод дифонного DTW-распознавания отдельно произносимых слов (в дальнейшем в этой статье мы будем говорить именно о распознавания отдельно произносимых слов). В качестве основного

инструмента сегментации используется численный аналог полной вариации, вычисляемый для последовательных отрезков по 256 отсчетов:

$$V = \sum_{i=0}^{254} \left| x_{i+1} - x_i \right|.$$

Поскольку во всех наших системах такая сегментация выполняется сразу после записи, и предшествует всем процедурам распознавания, ее естественно называть априорной.

Распознавание речи на всех этапах, за исключением возможно автоматического транскрибирования слов распознаваемого словаря, связано со случайными процессами, что является основным источником возможных ошибок. Это относится и к априорной сегментации. Как отмечено в [2], в большинстве случаев ошибки сегментации не влияют на результат распознавания. Однако они становятся существенными в следующей ситуации. Если сказанное слово распознано ошибочно, то пользователь может ввести в соответствующее поле правильное слово, и программа будет знать имена дифонов базы, которые нужны для построения эталона этого слова. Если сегментация будет правильной, то можно правильно выделить и прозвучавшие дифоны. При этом важно, чтобы метки сегментации стояли в нужных местах, а идентификация отрезков разбиения в терминах широкой фонетической классификации W-C-F-P не существенна. В этом случае распознающую систему можно дообучить, усреднив дифоны сказанного слова и соответствующие дифоны базы. Использование модифицированных дифонов при создании эталонов слов словаря будет приводить к улучшению распознавания данного диктора.

В связи с этим возникает задача сегментации при условии указания сказанного слова – условной сегментации. Такое понятие введено в работе [3]. Мы предлагаем осуществлять условную сегментацию путем модификации априорной сегментации. Первое упоминание о соответствующих процедурах сдержится в работе [4]. В настоящей статье мы систематизируем упоминавшиеся и опишем некоторые новые алгоритмы.

1. Прежде всего, программа должна выяснить имеются или нет ошибки в априорной сегментации (в случае их отсутствия коррекция, естественно, не нужна). Далее, в случае наличия ошибок, программа должна обнаружить те места, где они сделаны. Для этого по введенному слову строится его транскрипция, а затем обобщенная транскрипция в терминах широкой фонетической классификации (ШФК).

Пусть для примера сказано слово «пальма» и для него получилась сегментация

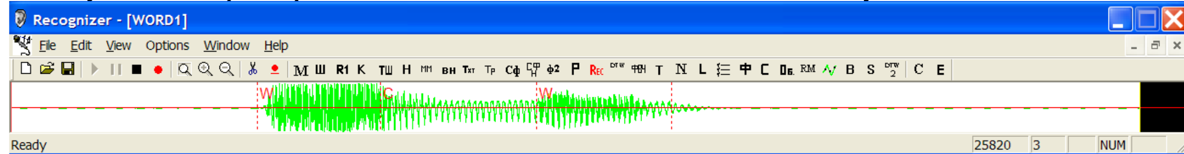


Рисунок 1 – Визуализация сигнала для слова «пальма» с ошибочной априорной сегментацией.

с сегментами W,C,W (1)

Наш автоматический транскриптор затранскрибирует слово так: «palma». Начальное *П* при априорной сегментации не выделяется. Поэтому использование ШФК-транскрипции выделяет гласный, два согласных и еще один гласный звук:

WCCW (2)

Программа последовательно сопоставляет символы (2) и символы (1), идя слева направо. Она обнаруживает, что второму символу C последовательности (2) соответствует W цепочки (1). Это указывает наличие и место ошибки. Коррекция будет состоять в делении сегмента C пополам с помощью дополнительной C-метки:

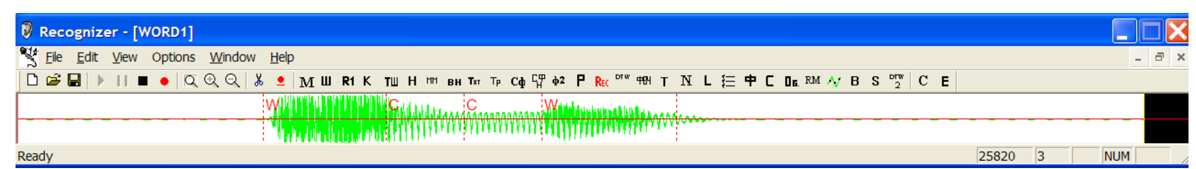


Рисунок 2 – Сегментация сигнала для слова «пальма» после коррекции

Аналогично осуществляется анализ и коррекция при обнаружении в априорной сегментации двух рядом стоящих С-сегментов вместо одного (вторая С-метка убирается). В случае, когда для слова «пальма» получается априорная сегментация с немаркированной меткой (рис. 3), последняя заменяется маркированной С-меткой.

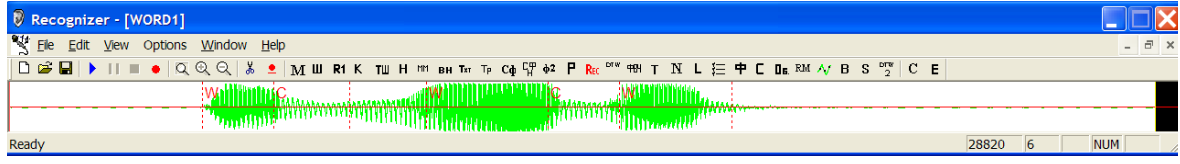


Рисунок 3 – Априорная сегментация для слова «пальма» с немаркированной меткой

Аналогично осуществляется коррекция сегментации на участках глухих звуков. Отметим, что для *Ц* и *Ч* считаются правильными две сегментации: PF и F.

2. На рис. 4 представлена визуализация сигнала для слова «облом» с априорной сегментацией, содержащей гласную вставку между звуками [б] и [л].

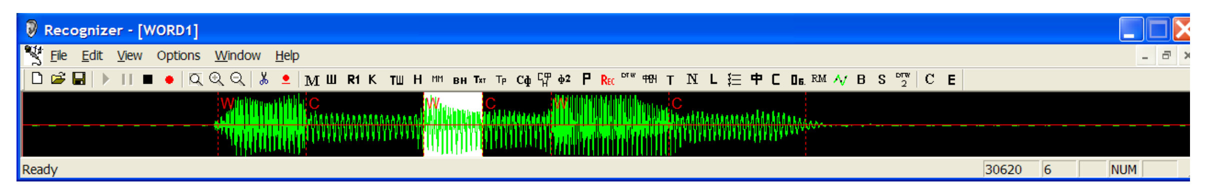


Рисунок 4 – Визуализация сигнала для слова «облом» с гласной вставкой между [б] и [л]

Программа обнаруживает ее описанным выше способом, сопоставляя цепочку

W,C,W,C,W,C

и ШФК-транскрипцию WCCWC. В результате коррекции средний W-сегмент убирается (вторая С-метка становится на место второй W-метки).

3. Встречаются случаи, когда для слова, начинающегося звонким согласным, априорная сегментация ошибочно ставит в начале лишний гласный звук. Чаще всего это случается для слов, начинающихся на [з], пример на рис. 5:

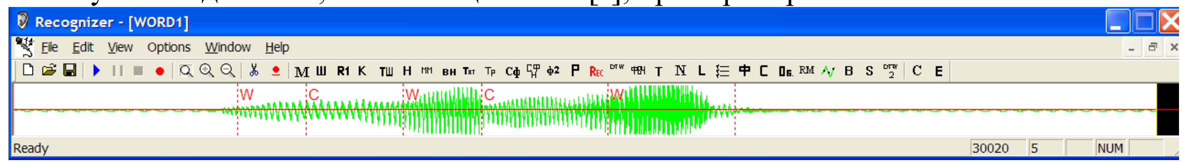


Рисунок 5 – Сегментация сигнала для слова «зима» с ошибочным гласным в начале

При коррекции маркировка начальной метки заменяется на С, а следующая С-метка убирается. Результат на рис. 6:

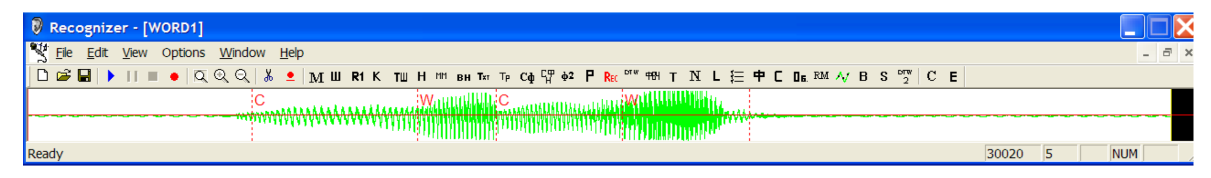


Рисунок 6 – Результат коррекции сегментации для слова «зима»

4. Остановимся на случае, когда при априорной сегментации не выделен С-сегмент перед глухим звуком (рис. 7). Наличие и местоположение этой ошибки определяется так же, как и выше. Коррекция осуществляется путем отдельной сегментации соответствующего W-отрезка; результат представлен на рис. 8. В случае, если дополнительная метка при этом все же не появляется, используется искусственное разбиение отрезка «равномерными» метками: он делится на 3 равные части и последняя треть считается искомым С-сегментом.

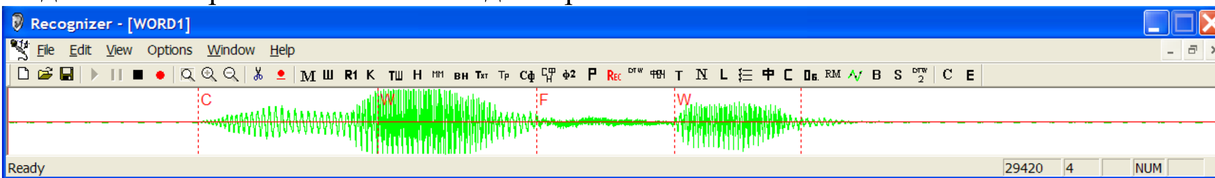


Рисунок 7 – Визуализация сигнала для слова «больше» с ошибочной априорной сегментацией

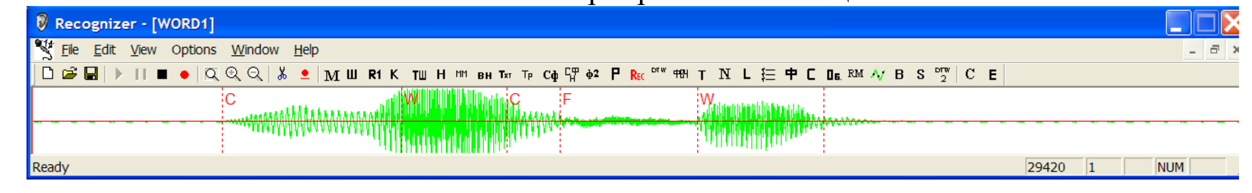


Рисунок 8 – Сегментация предыдущего сигнала после коррекции

Аналогично исчерпывается случай, когда при априорной сегментации не выделен С-сегмент после глухого звука.

В случае, когда слово заканчивается глухим звуком, но соответствующий заключительный отрезок в сегментации отсутствует, последний добавляется искусственно:

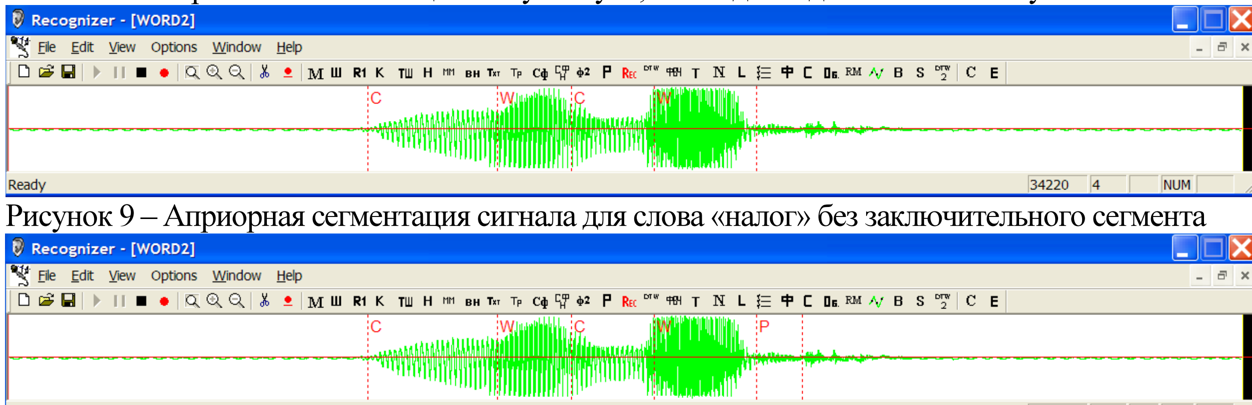


Рисунок 9 – Априорная сегментация сигнала для слова «налог» без заключительного сегмента

Рисунок 10 – Сегментация предыдущего сигнала после коррекции.

5. Обсудим теперь достаточно частую ошибку, когда при априорной сегметации не выделяется твердый или мягкий звук [р]. Это бывает, когда этот звук произносится не раскатисто, с неактивной артикуляцией. Здесь рассмотрим два отдельных случая.

а) Звук [р] находится между двумя гласными и при сегментации его следует искать внутри самого длинного W-отрезка.

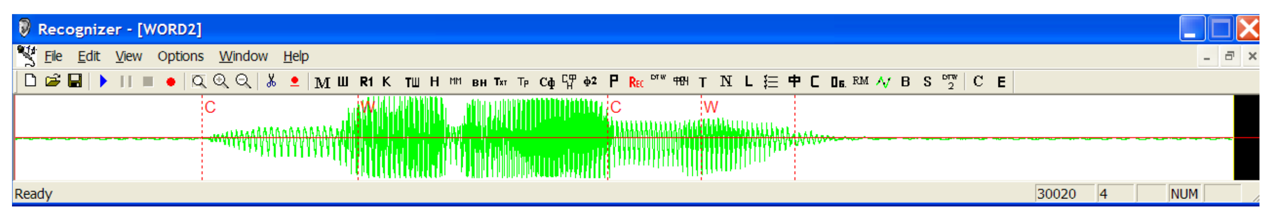


Рисунок 11 – Априорная сегментация для слова «ворона» с отсутствующим сегментом [р]

В этом случае этот W-отрезок разбивается «равномерными» метками на 3 равные части и средняя треть выделяется как отрезок звука [р] (рис. 12).

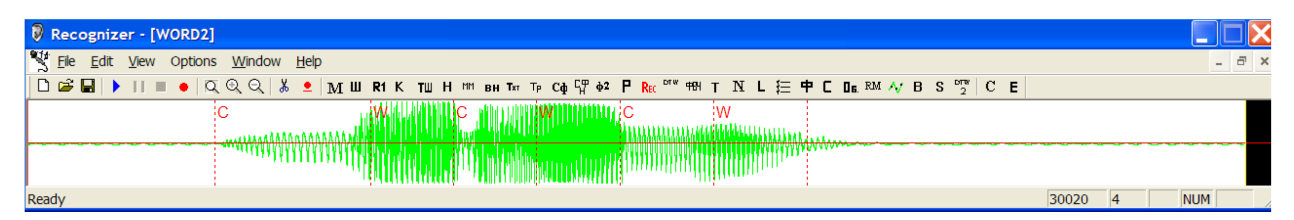


Рисунок 12 – Сегментация предыдущего сигнала после коррекции

b) Звук [р] предшествует звонкому согласному (рис. 13)

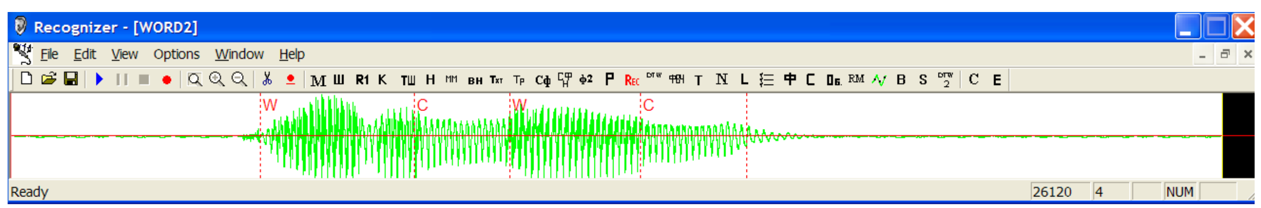


Рисунок 13 – Априорная сегментация для слова «карман» с отсутствующим сегментом [р]

В этом случае мы с помощью алгоритма, предложенного в [5], выделяем участки, соответствующие ударам языка о нёбо (р-удары, рис. 14) и отдельно сегментируем отрезок от первой р-метки до конца следующего С-отрезка. Результат представлен на рис. 15

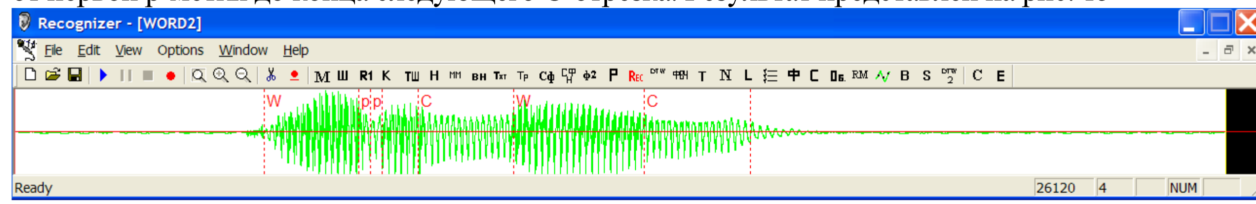


Рисунок 14 – Сегментация слова «карман» с выделением р-удара

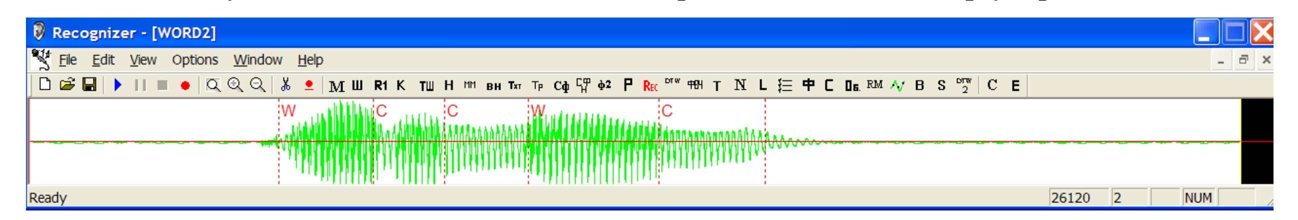


Рисунок 15 – Результат коррекции сегментации для слова «карман»

c) Звук [р] следует за звонким согласным. Этот случай исчерпывается аналогично предыдущему, только движение происходит не слева направо, а справа налево.

d) Звук [р] соседствует с глухим звуком. Здесь алгоритм коррекции такой же, как в случаях b) и c) с заменой звонкого согласного на глухой.

6. Случай двух рядом стоящих гласных.

Если участок, где находятся такие гласные, выделен, то желаемая метка, которая будет центром соответствующего дифона, может быть получена следующим образом. Участок разбивается «равномерными» метками на 3 равные части, средняя треть удаляется, а оставшиеся две части всего сигнала склеиваются. Таким образом, в данном случае происходит не только коррекция сегментации, но и преобразование распознаваемого сигнала. Однако это преобразование находится в русле того, что мы делаем, когда производим при распознавании слова межфонемную обработку [2]. При этом мы

получаем дифон, который можно путем усреднения с соответствующим дифоном дифонной базы использовать для модификации последнего в интересах распознавания для данного диктора.

Выделение участка, где находятся два соседних гласных, представляет наиболее трудную часть обсуждаемой проблемы, ибо при априорной сегментации он первоначально может, как выделяться целиком, так и разбиваться на два или три разноименных отрезка. Например, участок звукосочетания *АИ* стабильно сегментируется как WC. Добиться желаемого выделения можно, если процедуру типа той, которая описывалась выше относительно последовательностей (1) и (2) провести, двигаясь не только слева направо, но и справа налево.

Список литературы

1. Шелепов В.Ю. Построение системы голосового управления компьютером на примере задачи набора математических формул / В.Ю. Шелепов, А.В. Жук, А.В. Ниценко // Искусственный интеллект. – 2010. – № 3. – С.259-267.
2. Сегментация и дифонное распознавание речевых сигналов / А.К. Бурибаева, Г.В. Дорохина, А.В. Ниценко, В.Ю. Шелепов // Тр. СПИИРАН. – 31 (2013). – С. 20-42.
3. Козлов А.В. Система пофонемного распознавания отдельно произносимых слов / Козлов А.В., Саввина Г.В., Шелепов В.Ю. // Искусственный интеллект. №1, 2003. С.156-165.
4. Шелепов В.Ю. О некоторых вопросах, связанных с дифонным распознаванием и распознаванием слитной речи / Шелепов В.Ю., Ниценко А.В., Дорохина Г.В. // Искусственный интеллект. – 2013. – №3 – С. 209-216.
5. Шелепов В.Ю Обнаружение и выделение звука [р] в речевом сигнале / В.Ю. Шелепов, М.Х. Карабалаева, А.В. Ниценко // Искусственный интеллект. – 2011. – № 1. – С. 168-174.

References

1. Shelepov V.Ju. Postroenie sistemy golosovogo upravlenija komp'juterom na primere zadachi nabora matematicheskih formul / V.Ju. Shelepov, A.V. Zhuk, A.V. Nicenko // Iskusstvennyj intellekt. – 2010. – № 3. – S.259-267.
2. Segmentacija i difonnoe raspoznavanie rechevyh signalov / A.K. Buribaeva, G.V. Dorohina, A.V. Nicenko, V.Ju. Shelepov // Tr. SPIIRAN. – 31 (2013). – S. 20-42.
3. Kozlov A.V. Sistema pofonemnogo raspoznavanija otdel'no proiznosimyh slov / Kozlov A.V., Savvina G.V., Shelepov V.Ju. // Iskusstvennyj intellekt. №1, 2003. S.156-165.
4. Shelepov V.Ju. O nekotoryh voprosah, svjazannyh s difonnym raspoznavaniem i raspoznavaniem slitnoj rechi / Shelepov V.Ju., Nicenko A.V., Dorohina G.V. // Iskusstvennyj intellekt. – 2013. – №3 – C. 209-216.
5. Shelepov V.Ju Obnaruzhenie i vydelenie zvuka [r] v rechevom signale / V.Ju. Shelepov, M.H. Karabalaeva, A.V. Nicenko // Iskusstvennyj intellekt. – 2011. – № 1. – S. 168-174.

RESUME

.Ju. Shelepov, V A.V. Nicenko

Segmentation of Speech Signal which Corresponds to Beforehand Known Word

The article is devoted to segmentation of speech signal of the beforehand known word. Authors a priory segmentation for any speech signal acts as the basis. Generalized transcription within the framework of wide phonetic classification is controlling information. Algorithms are proposed for adding wanting and elimination of unnecessary segments of vowels, voiced and unvoiced consonants. Separation of hard and soft sound [r] and work with signal containing two neighboring sounds are subjects of the special attention. Results play a certain part in the modification diphone-base for concrete speaker in the process of diphone DTW-recognition.

Статья поступила в редакцию 05.06.2014.